A Predictive Study of Dirichlet Process Mixture Models for Curve Fitting

A Predictive Study of Dirichlet Process Mixture Models for Curve Fitting

SARA WADE

Computational and Biological Learning Laboratory, University of Cambridge

STEPHEN G. WALKER

Department of Mathematics and Division of Statistics and Scientific Computation, University of Texas at Austin

SONIA PETRONE

Department of Decision Sciences, Bocconi University

ABSTRACT. This paper examines the use of Dirichlet process (DP) mixtures for curve fitting. An important modelling aspect in this setting is the choice between constant or covariate-dependent weights. By examining the problem of curve fitting from a predictive perspective, we show the advantages of using covariate-dependent weights. These advantages are a result of the incorporation of covariate proximity in the latent partition. However, closer examination of the partition yields further complications, which arise from the vast number of total partitions. To overcome this, we propose to modify the probability law of the random partition to strictly enforce the notion of covariate proximity, while still maintaining certain properties of the DP. This allows the distribution of the partition to depend on the covariate in a simple manner and greatly reduces the total number of possible partitions, resulting in improved curve fitting and faster computations. Numerical illustrations are presented.

Key words: Dirichlet process, mixture models, random partitions, prediction

1 Introduction

Bayesian nonparametric curve fitting is an important area of research. The basic model is of the type

$$Y_i = m(x_i) + \sigma(x_i)\,\varepsilon_i, \tag{1}$$

where the curve $m(\cdot)$ is the focus of attention. Here $\sigma(\cdot)$ is a variance function and the (ε_i) are typically assumed to be independent and standard normal errors. Several methods have been developed in the literature; we refer to Denison *et al.* (2002, chap. 3) for an overview with focus on approaches using basis functions, and to Rasmussen & Williams (2006) for methods based on Gaussian processes. Further and more recent proposals can be found in DiMatteo *et al.* (2001) and Fan *et al.* (2010).

The Bayesian approach to curve fitting consists of assigning a prior on the random function $m(\cdot)$ and combining this prior with model (1) to compute the posterior given the data $(x, y) \equiv ((x_i, y_i), i = 1, \ldots, n)$. Then, the Bayesian curve estimate at x_0, with respect to the quadratic loss, is $\hat{m}(x_0) = \mathrm{E}[m(x_0)|x, y, x_0]$. It is worth to underline that $\hat{m}(x_0)$ corresponds to the point prediction, with respect to the quadratic loss, of the response at x_0, $\widehat{Y}(x_0) = \mathrm{E}[Y|x, y, x_0]$. Thus, examining predictive properties of flexible regression models provides another approach to solving the curve fitting problem. This is the approach that we adopt in this paper.

Mixture models based on the Dirichlet process (DP) are becoming an increasing popular tool for flexible regression, due to their ability to approximate a large class of conditional densities and their attractive balance between smoothness and flexibility in modelling local features. The general aim of this paper is to examine in detail properties of DP mixture models for curve fitting or, equivalently, their predictive properties.

The general form of a DP Gaussian mixture model for regression can be expressed as

$$Y|x, w, \mu, \sigma^2 \overset{\text{ind}}{\sim} \sum_{j=1}^{\infty} w_j(x)\mathrm{N}(\mu_j(x), \sigma_j^2(x)), \tag{2}$$

where $\mathrm{N}(a, b)$ denotes a normal distribution with mean a and variance b; and (w, μ, σ^2) denotes the collection of weight, mean, and variance functions of x, such that for each x, $\sum_{j=1}^{\infty} w_j(x) = 1$. Of course, in curve fitting, x is non-random. Thus, the above is not necessarily a conditional distribution, but the conditioning is a convenient notation. Model (2) implies that the choice of $m(\cdot)$ is given by

$$m(x) = \mathrm{E}[Y|x, w, \mu, \sigma^2] = \sum_{j=1}^{\infty} w_j(x)\mu_j(x). \tag{3}$$

Instead of having a "simple" distribution about this mean, which is usually assumed to be normal, model (2) allows flexible error distributions.

The key differences distinguishing the various proposals of form (2) present in the literature are in the construction and prior for the weight, mean, and variance functions. The Dirichlet

process mixture of linear regression models (DPM) is one of the earliest and simplest proposals. It assumes that the weights do not depend on x, and, within each mixture component, the variance is constant and the mean function is linear, $\mu_j(x) = \beta_j' x$. An early overview of Dirichlet process mixtures of linear models, with applications, is the article by West *et al.* (1994). The development of a software package in **R** (Jara (2007)) has eased the computational difficulties of implementing the model, and hence additionally increased the popularity of the model.

Müller *et al.* (1996) were the first to propose modelling the joint distribution of the dependent and independent variables as DP mixture of multivariate normals in order to obtain inference on the distribution of $Y|x$. For this model, again, the variance does not depend on x and the mean function has a linear form within cluster. However, the weights do depend on x. Further developments of this model can be found in Kang & Ghosal (2009), Shahbaba & Neal (2009), Hannah *et al.* (2011), Park & Dunson (2010), and Müller & Quintana (2010). Of course, this approach assumes that both x and y are random, even if the focus is on estimating $m(x)$.

MacEachern (1999) gave a general framework for nonparametric regression through models of the form (2) using dependent Dirichlet processes (DDP). Model (2) is regarded as a mixture of Gaussians where marginally the mixing distribution, $P_x = \sum_{j=1}^{\infty} w_j(x)\delta_{(\mu_j(x),\sigma_j^2(x))}$, is a Dirichlet process, and dependence is introduced among the random distributions P_x for varying x; the notation δ_a denotes the Dirac measure which is a probability measure with mass one on the point a. It has been shown (MacEachern (2000), Barrientos *et al.* (2012), Pati *et al.* (2013), Norets & Pelenis (2012)) that desirable properties such as large support and posterior consistency are possessed by simpler constructions that assume constant weight, mean, or variance functions. Motivated by these results and the desire for simple computations, many authors have focused on *single-p* DDPs, which assume constant weight functions. Usually, the variance function is also assumed to be constant in x, and the mean function is given a Gaussian process prior (Gelfand *et al.* (2005)) or assumed to be a linear function of a transformation of x into a higher dimensional space, $\mu_j(x) = \beta_j'\phi(x)$ (De Iorio *et al.* (2004)). For $\phi(\cdot)$ equal to the identity transformation, $\mu_j(x) = \beta_j' x$, the single-p DDP with linear mean functions (De Iorio *et al.* (2009) and Jara *et al.* (2010)) corresponds to the DPM model. More generally, the weights may also vary with x. Proposals to allow for covariate dependent weights include Griffin & Steel (2006), Dunson & Park (2008), Ren *et al.* (2011), and Rodriguez & Dunson (2011), just to mention a few. In these approaches, the mean functions are typically assumed constant or linear in x.

It clearly appears from (3) that a crucial modelling aspect is the choice between constant

and covariate-dependent weights. Thus, the first step of our study is a comparison between models with constant and covariate-dependent weight functions, when the focus is curve fitting and prediction. In particular, we will compare the DPM, as the basic model of the form (2) with constant weight functions, and the joint DPM model, as the computationally simplest model with covariate-dependent weights.

The choice of the weight function is indeed crucial for the predictive performance of the model. The weight functions have implications on the latent partition of the data in different mixture components, and prediction is strongly dependent on such partition.

Models with constant weight functions implicitly assume that the covariates are not informative on the cluster allocation. This may be appropriate for exploratory analysis, aimed at highlighting possible clusterings of individual regression curves. However, in curve fitting, clustering is not meant to model heterogeneity, i.e. a multiple response behavior for the same region of x, but rather aims at possibly selecting different curves, from the collection of available curves $\mu_j(\cdot)$, in different regions of the covariate space, for local approximation of the unknown regression curve. In this context, we show that the assumption of a constant weight function can result in (surprisingly) poor and uninformative prediction, the more so in case of departures of the real curve from the form specified by the mean functions. As we will highlight later, this occurs because, for a given partition, the prediction is a mixture of all the cluster-specific fitted curves, independent of x_{n+1} and the location of the clusters in the covariate space.

Models with covariate dependent weights implicitly use a notion of covariate-proximity clustering that greatly improves prediction. For a given partition, predictions based on clusters which are close to x_{n+1} in the covariate space have greater influence, and the conditional predictions are then averaged across all partitions, according to the posterior distribution. Unfortunately, as we will illustrate, the information about what are reasonable, proximity-based partitions gets (dramatically) spread out in the posterior, leading to predictions based on undesirable partitions having too much impact and predictions based on desirable partitions with not enough impact.

These difficulties arise due to the huge number of partitions on which DP-based models assign a prior distribution. In particular, both models allow for any possible partition of the n data points into k groups for $k = 1, \ldots, n$. There are

$$S_{n,k} = \frac{1}{k!} \sum_{j=0}^{k} (-1)^j \left(\begin{array}{c} k \\ j \end{array} \right) (k-j)^n,$$

a Stirling number of the second kind, ways to partition the n data points in to the k groups and

$$B_n = \sum_{k=1}^{n} S_{n,k},$$

a Bell number, possible partitions of the n data points. Even for small n, this number is very large.

However, the covariates typically provide information on the partition structure. Our main point is that this information should be strictly enforced in the prior probability law on the random partition, since it would otherwise be (dramatically) spread out in the posterior, due to the huge dimension of the partition space. In particular, if we require partitions to satisfy an ordering constraint of the (x_i), we can reduce the total number of partitions to just 2^{n-1} of the B_n total partitions. For example, for $n = 10$, the total number of partitions under this constraint is 0.44% of the total partitions, and for $n = 100$ the percentage of partitions under this constraint is less than 10^{-83}% of the total partitions. Clearly, this set of desirable partitions is much smaller than the partition space, and thus, defining a prior on the partition that ensures sufficient mass on the desirable partitions in the posterior can be difficult.

To resolve this issue, we propose to modify the distribution of the latent partition to rule out the undesirable partitions by setting the probability of these events to be zero, while still maintaining properties of the DP, such as the prior for k_n, the number of groups in a sample of size n. This allows the distribution of the partition to depend on the covariate according to the designated clustering principle and greatly reduces the number of possible partitions. Our aim is to demonstrate greatly improved prediction. Furthermore, due to the reduced dimension of the partition space, computations are much less expensive.

The research in this paper is motivated by a data set consisting of possible Alzheimer's disease (AD) patients with measurements of the volume of different brain structures. The interest is in estimation of the curve describing the probability of AD as a function of asymmetry of the hippocampus. Nonparametric flexibility is needed to recover the non-monotone curve.

The paper is organized as follows. In Section 2 we review the DPM and joint DPM models, the implied random partition models, and the prediction under the models. In Section 3 we recalibrate the DPM to remove undesirable partitions and obtain useful posterior and predictive distributions. Section 4 covers the computational procedures for sampling and prediction under the restricted DPM model. Finally, numerical illustrations are presented in Section 5 and an AD study is presented in Section 6.

2 DPM and joint DPM models

2.1 DPM model

The Dirichlet process prior defines a probability law on distributions on arbitrary spaces and was first introduced by Ferguson (1973). Mixtures models with a Dirichlet process mixing distribution, of the type we will be using, were subsequently introduced and studied by Lo (1984). The DP mixture model for the distribution of response, Y_i, given the covariate, x_i, for $i = 1, \ldots, n$, has the form

$$\begin{aligned} Y_i|x_i, \beta_i, \sigma_i^2 &\overset{\text{ind}}{\sim} \mathrm{N}(\beta_i' \underline{x}_i, \sigma_i^2), \\ (\beta_i, \sigma_i^2)|P &\overset{\text{iid}}{\sim} P, \\ P &\sim \mathrm{DP}(\alpha P_0), \end{aligned} \tag{4}$$

where $\underline{x}_i = (1, x_i)'$ and, for convenience, we condition on x_i even when the covariate is non-random. Here, the base measure, P_0, is the conjugate multivariate normal–inverse gamma distribution, i.e $\beta|\sigma^2 \sim \mathrm{N}(\beta_0, \sigma^2 C^{-1})$ and $\sigma^2 \sim \mathrm{IG}(a, b)$, for some selection of (β_0, C, a, b).

From the properties of the DP, P is discrete with probability one, implying positive probabilities of ties among the parameters pairs $(\beta_i, \sigma_i^2)_{i=1}^n$. This follows from the structure of the predictive distributions, which is given by the Pólya urn scheme (Blackwell & MacQueen (1973))

$$\begin{aligned} (\beta_1, \sigma_1^2) &\sim P_0, \\ (\beta_{n+1}, \sigma_{n+1}^2)|(\beta_1, \sigma_1^2), \ldots, (\beta_n, \sigma_n^2) &\sim \frac{\alpha}{\alpha + n} P_0 + \sum_{j=1}^{k_n} \frac{n_{n,j}}{\alpha + n} \delta_{(\beta_j^*, \sigma_j^{2*})}, \end{aligned}$$

where $(\beta_1^*, \sigma_1^{2*}), \cdots, (\beta_{k_n}^*, \sigma_{k_n}^{2*})$ are the k_n distinct values in the sample $(\beta_1, \sigma_1^2), \ldots, (\beta_n, \sigma_n^2)$, in order of appearance, and $n_{n,j} = \sum_{i=1}^n I_{(\beta_i, \sigma_i^2) = (\beta_j^*, \sigma_j^{2*})}$ are their frequencies. For ease of notation, we drop the subscript n from $(k_n, n_{n,j})$ when the sample size is understood.

The DPM model can be equivalently viewed in terms of a random partition model that gives the distribution of the partition of n subjects into clusters (**?**), and a sampling model, which models the data given the partition. Let $\rho_n = (s_1, \ldots, s_n)$ denote the partition, where $s_i = j$ if $(\beta_i, \sigma_i^2) = (\beta_j^*, \sigma_j^{2*})$. The random partition model is obtained from the Pólya urn scheme

$$p(\rho_n) = \frac{\alpha^k}{\alpha^{[n]}} \prod_{j=1}^{k} (n_j - 1)!, \tag{5}$$

where $\alpha^{[n]} = \alpha(\alpha+1)\cdots(\alpha+n-1)$. From (4), the sampling model for the response given the partition and the covariate assumes independence across clusters and exchangeability within cluster, where conditional on the cluster parameters, a simple linear model is assumed within cluster.

Note that the partition of the n observations is independent of x. This means that given the covariates, positive mass is assigned to any possible partition of the n observations into k groups and that there is no prior preference for clusters with similar covariates.

The posterior of the partition given the observed data, $((y_i, x_i), i = 1, \ldots, n)$, denoted (y, x) for brevity, is proportional to the random partition model, times the sampling model. The use of conjugate base measures in (4) allows for a closed form expression for the sampling model and combining this expression with the prior, implies the posterior of a partition is

$$p(\rho_n|y,x) \propto \alpha^k \prod_{j=1}^{k}(n_j-1)! \left(\frac{|C|}{|C+X_j'X_j|}\right)^{1/2} \frac{b^a\Gamma(a+n_j/2)}{\Gamma(a)(b+V_j^2/2)^{a+n_j/2}}, \tag{6}$$

where

$$V_j^2 = (\underline{y}_j - \widehat{\underline{y}}_j)'\widehat{W}_j(\underline{y}_j - \widehat{\underline{y}}_j);$$

$$\widehat{W}_j = (I - X_j(C+X_j'X_j)^{-1}X_j');$$

$$\widehat{\underline{y}}_j = X_j\beta_0;$$

$\underline{y}_j$ denotes the response of data points in cluster j; and X_j is a matrix whose rows consist of $\underline{x}_i$ for data points in cluster j. Equation (6) shows that partitions with similar linear relationships between y and x are preferred in the posterior.

Due to the large number of possible partitions, direct computation of (6) is unfeasible and requires MCMC approximations. We let $l = 1, \ldots, L$ index the iterations of a MCMC output, $\{\rho_n^{(l)}\}_{l=1}^L$, where for each l, $\rho_n^{(l)}$ is an approximate sample from the posterior distribution of $[\rho_n|y,x]$. Due to the huge dimension of the partition space, the chain will tend to visit too many partitions with each one only visited very few times.

Under quadratic loss, the curve estimate at x_{n+1} corresponds to the point prediction of Y at x_{n+1}:

$$\widehat{m}(x_{n+1}) = \mathrm{E}[Y_{n+1}|x_{n+1}, y, x].$$

Let $\mathcal{P}_n$ denote the set of all partitions of $\{1, \ldots, n\}$ and $\mathcal{P}(\rho_n) = \{1, ..., k+1\}$ denote the possible labels for the new data point given ρ_n; then, since the prior on the random partition does not

depend on the covariates,

$$\widehat{m}(x_{n+1}) = \sum_{\rho_n \in \mathcal{P}_n} \left(\sum_{s_{n+1} \in \mathcal{P}(\rho_n)} \mathrm{E}[Y_{n+1}|x_{n+1}, y, x, \rho_{n+1}] p(s_{n+1}|\rho_n) \right) p(\rho_n|y, x). \tag{7}$$

The inner term of (7), the prediction given ρ_n, is simply an average of all cluster-specific predictions with weights given by the Pólya urn scheme;

$$\mathrm{E}[Y_{n+1}|x_{n+1}, \rho_n, x, y] = \frac{\alpha}{\alpha + n} \beta_0' \underline{x}_{n+1} + \sum_{j=1}^{k} \frac{n_j}{\alpha + n} \widehat{\beta}_j' \underline{x}_{n+1}, \tag{8}$$

where

$$\widehat{\beta}_j = (C + X_j' X_j)^{-1} (C\beta_0 + X_j' \underline{y}_j)$$

is a vector containing the estimated intercept and slope for the regression line under the standard linear model given the response and covariates of subjects in cluster j.

Equation (8) shows that given the partition, the cluster-specific predictions are weighted according to the size of each cluster. This means that even if the new x_{n+1} is very far from the largest group, it is more likely to share the same regression line because many observations fall in that group. This aspect can clearly lead to very poor curve fitting and prediction.

Using equation (8), the expression for the curve estimate given in (7) becomes

$$\widehat{m}(x_{n+1}) = \sum_{\rho_n \in \mathcal{P}_n} \left(\frac{\alpha}{\alpha + n} \beta_0' \underline{x}_{n+1} + \sum_{j=1}^{k} \frac{n_j}{\alpha + n} \widehat{\beta}_j' \underline{x}_{n+1} \right) p(\rho_n|x, y),$$

which can be approximated through MCMC by

$$\widehat{m}(x_{n+1}) \approx \frac{1}{L} \sum_{l=1}^{L} \left(\frac{\alpha}{\alpha + n} \beta_0' \underline{x}_{n+1} + \sum_{j=1}^{k^{(l)}} \frac{n_j^{(l)}}{\alpha + n} \widehat{\beta}_j'^{(l)} \underline{x}_{n+1} \right). \tag{9}$$

Thus, the prediction is averaged across all partitions, with weights given by their (estimated) posterior probability, and will therefore suffer from the issues for the posterior of the partition, namely the insufficiently large posterior mass of desirable partitions and insufficiently small posterior mass of undesirable partitions. If the prediction is based on an undesirable partition, the estimated regression line and/or weights within cluster be will be incorrect and the poor prediction resulting from this undesirable partition will be used in computations of (9). These issues are illustrated with examples in Section 5.

Also note that factoring out the $\underline{x}_{n+1}$ yields

$$\widehat{m}(x_{n+1}) = \left(\frac{\alpha}{\alpha + n} \beta_0 + \sum_{\rho_n \in \mathcal{P}_n} \sum_{j=1}^{k} p(\rho_n|x, y) \frac{n_j}{\alpha + n} \widehat{\beta}_j \right)' \underline{x}_{n+1}.$$

Thus, the curve estimate is merely a linear function of x_{n+1}, meaning that no matter where x_{n+1} lies in the covariate space, the same linear function is used to estimate y_{n+1}.

2.2 Joint DPM model

The joint DPM model is similar to (4), but also incorporates a model for the covariate,

$$\begin{aligned} Y_i|x_i, \beta_i, \sigma^2_{y,i} &\overset{\text{ind}}{\sim} \mathrm{N}(\beta_i' \underline{x}_i, \sigma^2_{y,i}), \\ X_i|\theta_i &\overset{\text{ind}}{\sim} F_X(\cdot;\theta_i), \\ (\beta_i, \sigma^2_{y,i}, \theta_i)|P &\overset{\text{iid}}{\sim} P, \\ P &\sim \mathrm{DP}(\alpha P_{0Y} \times P_{0X}), \end{aligned}$$

where P_{0Y} is the base measure for the Y parameters and P_{0X} is the base measure for the X parameters. We assume the same structure for P_{0Y}, namely, the conjugate multivariate normal–inverse gamma for some selection of (β_0, C, a, b), and do not assume a specific form for P_{0X}, but for the examples in Section 5, where F_X is the normal distribution function, it is chosen to be the conjugate normal–inverse gamma.

As for the DPM, also the joint DPM model can be decomposed into a random partition model and a sampling model given the partition. However, different from the DPM, the random partition of (β_i, σ_i^2) depends on the covariates (Park & Dunson (2010)) and is given by

$$p(\rho_n|x) \propto \alpha^k \prod_{j=1}^{k} (n_j - 1)! \int \prod_{\{i \in S_j\}} f_X(x_i|\theta) dP_{0X}(\theta), \tag{10}$$

where $S_j = \{i : s_i = j\}$ and f_X is the density of F_X.

Müller & Quintana (2010) independently constructed a similar model, but were motivated by directly modifying the cohesion term of the random partition model by a factor that favors clusters with similar covariates. More specifically, they suggested to modify the partition distribution (5) of the DP by introducing a *similarity function* $g(\cdot)$ as follows

$$p(\rho_n|x) \propto \alpha^k \prod_{j=1}^{k} (n_j - 1)!\, g(\{x_i\}_{i \in S_j}),$$

where $g(\cdot)$ captures the closeness of covariates, with large values indicating high similarity. Müller & Quintana (2010) show that if the similarity function satisfies invariance with respect to permutations of the covariates and scalability, i.e

$$\int g(\{x_i\}_{i \in S_j}, x) dx = g(\{x_i\}_{i \in S_j}),$$

then

$$g(\{x_i\}_{i\in S_j}) = \int \prod_{\{i\in S_j\}} f_X(x_i|\theta) dP_{0X}(\theta);$$

and thus, the covariate dependent random partition model is equivalent to that obtained in (10). Even though (10) still assigns positive mass to any possible partition of the n subjects into k groups, clusters with similar covariates are encouraged.

The posterior of the covariate dependent partition is

$$p(\rho_n|y,x) \propto \alpha^k \prod_{j=1}^{k} (n_j-1)! g(\{x_i\}_{i\in S_j}) \left(\frac{|C|}{|C+X_j'X_j|}\right)^{1/2} \frac{b^a\Gamma(a+n_j/2)}{\Gamma(a)(b+V_j^2/2)^{a+n_j/2}}.$$

Due to the incorporation of the similarity function, desirable partitions have higher posterior mass, and the MCMC chain visits more reasonable partitions. However, the total number of partitions has not changed; undesirable partitions still have positive prior mass, and incorporation of the similarity function may not be enough to ensure their posterior mass is sufficiently small. Furthermore, there will likely still be many partitions which fit the data, resulting in posterior mass diluted across many partitions.

The curve estimate, i.e. the prediction of Y_{n+1} given x_{n+1} and the data, is again computed as

$$\widehat{m}(x_{n+1}) = \sum_{\rho_n\in\mathcal{P}_n} \left(\sum_{s_{n+1}\in\mathcal{P}(\rho_n)} \mathrm{E}[Y_{n+1}|x_{n+1},y,x,\rho_n,s_{n+1}] p(s_{n+1}|\rho_n,x,x_{n+1}) \right) p(\rho_n|y,x,x_{n+1}).$$

However, contrary to the DPM model, the cluster allocation of the new observation now depends on the covariate and is given by

$$s_{n+1}|\rho_n,x,x_{n+1} \sim \frac{1}{p(x_{n+1}|x,\rho_n)} \left(g(x_{n+1})\frac{\alpha}{\alpha+n}\delta_{k+1} + \sum_{j=1}^{k} g(x_{n+1}|\{x_i\}_{i\in S_j})\frac{n_j}{\alpha+n}\delta_j \right),$$

where the weights of the Pólya urn scheme are modified by the cluster-specific predictive densities of x_{n+1}:

$$g(x_{n+1}|\{x_i)\}_{i\in S_j}) = \int f_X(x_{n+1}|\theta) dP_{0X}(\theta|\{x_i\}_{i\in S_j}).$$

Furthermore,

$$p(\rho_n|x,y,x_{n+1}) = \frac{p(x_{n+1}|x,\rho_n)}{p(x_{n+1}|x,y)} p(\rho_n|x,y)$$

is no longer equivalent to $p(\rho_n|x,y)$. The resulting expression of the curve estimate is

$$\widehat{m}(x_{n+1}) = \sum_{\rho_n\in\mathcal{P}_n} \left(\frac{\alpha}{c} g(x_{n+1})\ \beta_0'\underline{x}_{n+1} + \sum_{j=1}^{k} \frac{n_j}{c} g(x_{n+1}|\{x_i\}_{i\in S_j})\ \widehat{\beta}_j'\underline{x}_{n+1} \right) p(\rho_n|y,x), \qquad (11)$$

where $c = (\alpha + n)p(x_{n+1}|y, x)$. The inner term of (11) is again an average of all cluster-specific predictions, but the weights here depend on the closeness of x_{n+1} to the clusters in the covariate space, as measured by the cluster-specific predictive densities. Regression lines for clusters close to x_{n+1} are assigned more weight. However, regression lines for clusters far from x_{n+1} in the covariate space still have positive weight, resulting in unnecessary inclusion of poor predictions based on these clusters in the average computed in (11).

The curve estimate (11) can be approximated by MCMC as

$$\approx \frac{1}{L}\sum_{l=1}^{L}\left(\frac{\alpha}{\hat{c}}g(x_{n+1})\beta_0'\underline{x}_{n+1} + \sum_{j=1}^{k^{(l)}}\frac{n_j^{(l)}}{\hat{c}}g(x_{n+1}|\{x_i\}_{i\in S_j^{(l)}})\widehat{\beta}_j^{(l)\prime}\underline{x}_{n+1}\right), \tag{12}$$

where

$$\hat{c} = \left(\frac{1}{L}\sum_{l=1}^{L}\alpha g(x_{n+1}) + \sum_{j=1}^{k^{(l)}} n_j^{(l)} g(x_{n+1}|\{x_i\}_{i\in S_j^{(l)}})\right).$$

Again, the estimate obtained in (12) by averaging over all partitions visited by the chain will suffer from the issues for the posterior of the partition mentioned above and poor prediction arising from undesirable partitions with insufficiently small posterior mass.

Finally, note that the curve estimate is no longer a linear function of x_{n+1}, since the weights assigned to each regression line depend on x_{n+1}.

3 A restricted DPM model

Our proposal extends the DPM model (4), which written in terms of the random partition is

$$(Y_1, \ldots, Y_n) \mid x, \rho_n, (\beta_1^*, \sigma_1^{2*}), \ldots, (\beta_{k_n}^*, \sigma_{k_n}^{2*}) \sim \prod_{j=1}^{k_n}\prod_{\{i:s_i=j\}} N(y_i|\beta_j^{*'}\underline{x}_i, \sigma_j^{2*})$$

$$\rho_n \sim p(\rho_n)$$

and $(\beta_j^*, \sigma_j^{2*}) \overset{\text{iid}}{\sim} P_0$. In the DPM model, the random partition model $p(\rho_n)$ is induced by the assumption of exchangeability of the individual parameters (β_i, σ_i^2) with a DP prior on their distribution. Here, as for the joint DPM model, we relax the assumption of exchangeability of the individual parameters to allow for a cluster allocation that is covariate dependent. Our proposal is a new random partition model $p^*(\rho_n|x)$ that strictly incorporates a covariate-proximity constraint.

Indeed, in regression settings where the covariate is informative for prediction, partitioning should be based on the proximity of the covariates. Due to the unrestricted nature of the

clusters offered by Dirichlet based models, this idea of covariate proximity needs to be specifically enforced on the partition structure.

For curve-fitting, the idea of covariate proximity is naturally expressed by the ordering of x. For example, if $x_i < x_{i'} < x_{i''}$, it is reasonable to assume that if subjects (i, i'') are clustered together, then subject i' is also in that cluster. To this aim, we use the natural ordering of x to determine the allowed partitions and remove undesirable partitions by adjusting the conditional distribution of partition given the covariate, so that their mass is zero.

Let π_x denote the permutation of the first n integers that rearranges $(x_1 \ldots, x_n)$ in increasing order, as $x_{\pi_x(1)} < \ldots < x_{\pi_x(n)}$, and let $y_{\pi_x(1)}, \ldots, y_{\pi_x(n)}$ and $s_{\pi_x(1)}, \ldots, s_{\pi_x(n)}$ be the corresponding values of y and $s_1, \ldots, s_n$. For the DP, the prior distribution (5) of the partition is invariant to a relabelling of the clusters as long as the partition is preserved. This means that we can relabel the clusters, so that the subject with the smallest covariate is in the first cluster. To impose the order constraint that if subjects i and i'' are clustered together then all subjects whose covariates are between x_i and $x_{i''}$ are in the same cluster, we require that

$$s_{\pi_x(1)} \leq \ldots \leq s_{\pi_x(n)}. \tag{13}$$

A similar constraint is enforced in Fuentes-Garcia *et al.* (2010); however, in their work, no covariates are present and the imposed restriction is based on the ordering of the observed data y, with the aim of improved inference on the clustering structure. They incorporate the restriction by simply multiplying the posterior of ρ_n by the indicator that the constraint is satisfied.

We first note that while a simple extension of their approach in a regression setting, i.e. multiplying $p(\rho_n|x)$ by the indicator that $s_{\pi_x(1)} \leq \ldots \leq s_{\pi_x(n)}$, does remove the unwanted partitions, it also leads to an undesirable prior for k. Indeed, such an approach would cause the prior for k to place a high mass on $k = 1$ and $k = n$, and for a fixed value of α, the mass assigned to $k = 1$ increases with the sample size. This unbalance effect is due to the fact that we are removing no partitions for $k = 1$ and $k = n$ and many as $k \to n/2$. The mass of the removed partitions is spread out evenly among the remaining partitions, thus increasing the relative weight of $k = 1$ and $k = n$, and decreasing the relative weight of moderate values of k.

To avoid this effect, we define a covariate dependent random partition model that both removes undesirable partitions and retains certain properties of the random partition model induced by the DP. More specifically, we want to modify the partition probability law (5) of the DPM model, but to keep unchanged the probability law of the frequencies $(m_1, \ldots, m_n)$

corresponding to cluster sizes $(n_1, \ldots, n_k)$, where m_j is the number of $n_1, \ldots, n_k$ that are equal to j. For the DP, the probability law of $(m_1, \ldots, m_n)$ is given by the celebrated Ewens sampling formula (Ewens (1972)). In addition, preserving the law of $(m_1, \ldots, m_n)$ implies that the probability law of the number of clusters k is unchanged. Our proposal is given in the following proposition.

Proposition 1. *The covariate-dependent probability measure on the random partition defined by*

$$p^*(\rho_n|x) = \frac{\alpha^k}{\alpha^{[n]}} \frac{n!}{k!} \prod_{j=1}^{k} \frac{1}{n_j} * I_{s_{\pi_x(1)} \leq \ldots \leq s_{\pi_x(n)}} \tag{14}$$

satisfies the order constraint (13) and has the same marginal for $(m_1, \ldots, m_n)$ and for k, as those induced by the Dirichlet process.

Proof. By construction, the random partition model p^* satisfies the order constraint (13). We want to show that it preserves the probability law of $(m_1, \ldots, m_n)$ induced by the DP. The proof relies on the fact that under constraint (13), the partition is uniquely identified by $(n_1, \ldots, n_k, k)$, that is, there is only one partition that has cluster sizes $(n_1, \ldots, n_k)$ and satisfies (13), namely $(s_{\pi_x(1)}, \ldots, s_{\pi_x(n)}) = (1, \ldots, 1, 2, \ldots, 2, \ldots, k, \ldots, k)$, where 1 is repeated n_1 times, 2 is repeated n_2 times, ..., and k is repeated n_k times. Therefore,

$$p^*(n_1, \ldots, n_k, k|x) = \frac{\alpha^k}{\alpha^{[n]}} \frac{n!}{k!} \prod_{j=1}^{k} \frac{1}{n_j}. \tag{15}$$

Now, $p^*(m_1, \ldots, m_n|x) = \sum_{\mathcal{N}} p^*(n_1, \ldots, n_k, k|x)$, where the sum is over all $(n_{\pi(1)}, \ldots, n_{\pi(k)})$ obtained from a permutation π of the clustering indices of a specific $(n_1, \ldots, n_k)$ that satisfies $m_1, \ldots, m_n$. Since (15) is invariant to a permutation of cluster indices, the probability of $(m_1, \ldots, m_n)$ is simply the probability of a specific $(n_1, \ldots, n_k)$ that satisfies $(m_1, \ldots, m_n)$ multiplied by the number of unique ways to order the m_i clusters of size i for $i = 1, \ldots, n$, that is

$$p^*(m_1, \ldots, m_n|x) = p^*(n_1, \ldots, n_k, k|x) \frac{k!}{\prod_{i=1}^{n} m_i!}.$$

This implies that

$$p^*(m_1, \ldots, m_n|x) = \frac{\alpha^k}{\alpha^{[n]}} \frac{n!}{k!} \prod_{j=1}^{k} \frac{1}{n_j} \frac{k!}{\prod_{i=1}^{n} m_i!} = \frac{\alpha^k}{\alpha^{[n]}} \frac{n!}{\prod_{i=1}^{n} i^{m_i} m_i!},$$

where the last step follows from noting that $n_1 \cdots n_k = 1^{m_1} 2^{m_2} \cdots n^{m_n}$. This is probability law of $(m_1, \ldots, m_n)$ induced by the DP (Antoniak (1974)). Notice that $k = \sum_{i=1}^{n} m_i$; thus, it follows that the prior for k is equivalent to that of the DP. $\diamond$

The proof relies on the fact that under constraint (13), there is only one partition described by $(n_1, \ldots, n_k, k)$. This property will also be exploited for computations in Section 4.

We note that the order-based dependent Dirichlet process models of Griffin & Steel (2006) also implicitly define a covariate dependent random partition model based on the ordering of covariate values. The important difference to underline is that we are not only encouraging order-based partitions, but also removing undesirable partitions which violate this constraint, greatly reducing the total number of partitions and ensuring undesirable partitions have zero posterior mass.

3.1 The posterior distribution

The posterior distribution of the partition is

$$p^*(\rho_n|y,x) \propto \frac{\alpha^k}{k!}\prod_{j=1}^{k}\frac{1}{n_j}\left(\frac{|C|}{|C+X_j'X_j|}\right)^{1/2} * \frac{b^a\Gamma(a+n_j/2)}{\Gamma(a)(b+V_j^2/2)^{a+n_j/2}} * I_{s_{\pi_x(1)}\leq\ldots\leq s_{\pi_x(n)}},$$

which depends on the hyper–parameters $(\alpha, C, \theta_0, b, a)$. The interpretation of these parameters is similar to the DP model. A large value for α will encourage more clusters through the factor of α^k. For a given k, the term $\prod_{j=1}^{k} n_j^{-1}$ will favour partitions with one large cluster and several small clusters. Thus, if one believes that the clusters are balanced, the prior distribution of the partition should be adjusted appropriately.

Given σ^2, the prior variance–covariance matrix of the intercept and slope is $\sigma^2 C^{-1}$. Typically, C is a diagonal matrix with small values on the diagonal so that the prior is non-informative. In this case, $|C| < 1$ and

$$\prod_{j=1}^{k}\left(\frac{|C|}{|C+X_j'X_j|}\right)^{1/2} \approx \frac{|C|^{k/2}}{\prod_{j=1}^{k}|X_j'X_j|^{1/2}}.$$

The term $|C|^{k/2}$ will discourage a large number of clusters, while

$$\prod_{j=1}^{k}|X_j'X_j|^{1/2} = \prod_{j=1}^{k} n_j\left(\sum_{i\in S_j}\frac{(x_i-\bar{x}_j)^2}{n_j}\right)^{1/2},$$

where $\bar{x}_j$ is the sample mean of the (x_i) in cluster j, will encourage clusters with similar values of the covariate and unbalanced clusters. For a given k, the term $\prod_{j=1}^{k}\Gamma(a+n_j/2)/\Gamma(a)$ will also encourage unbalanced clusters. Finally, $\prod_{j=1}^{k} b^a/(b+V_j^2/2)^{a+n_j/2}$ will encourage clusters with similar values of the covariate and similar linear response curve, since V_j^2 will be smaller in this case.

3.2 Prediction

Given the partition of the observed subjects and new subject, the predictive distribution has a known form and can be easily computed and sampled from. In particular, suppose that according to ρ_{n+1} the new subject is in cluster j. Then, the predictive distribution of Y_{n+1} is obtained from standard computations based on the observations in cluster j. In particular, it is a non-central t-distribution with location $\widehat{\beta}_j'\underline{x}_{n+1}$, scale $\widehat{b}_j^{-1}\widehat{a}_j\widehat{W}_{n+1,j}$, and $2a+n_j$ degrees of freedom:

$$(Y_{n+1}-\widehat{\beta}_j'\underline{x}_{n+1}) * \left(\frac{\widehat{a}_j\widehat{W}_{n+1,j}}{\widehat{b}_j}\right)^{1/2} |\rho_{n+1},y,x \sim \mathcal{T}(2a+n_j),$$

where $\mathcal{T}(\nu)$ denotes the t-distribution with ν degrees of freedom. Here we denote the response and covariate matrix for the n_j observed subjects in cluster j by $(X_j,\underline{y}_j)$; we define

$$\widehat{W}_{n+1,j} = 1-\underline{x}_{n+1}'(\widehat{C}_j+\underline{x}_{n+1}\underline{x}_{n+1}')^{-1}\underline{x}_{n+1},$$

$$\widehat{C}_j = C+X_j'X_j,$$

$$\widehat{a}_j = a+n_j/2, \text{ and } \widehat{b}_j = b+V_j^2/2,$$

and compute $\widehat{\beta}_j$ and V_j^2 based on $(X_j,\underline{y}_j)$. If the new subject belongs to a new cluster, then $n_j=0$ and the updated parameters, $\widehat{a}_j,\widehat{b}_j,\widehat{\beta}_j,\widehat{C}_j$ are given by the prior parameters.

Define $\mathcal{C}_n$ as the set of possible partitions of the n subjects under the restricted DPM model and $\mathcal{C}(\rho_n)$ as the set of values for s_{n+1} such that ρ_{n+1} restricted to n observed subjects is ρ_n. The curve estimate is again computed as

$$\widehat{m}(x_{n+1}) = \sum_{\rho_n\in\mathcal{C}_n}\left(\sum_{s_{n+1}\in\mathcal{C}(\rho_n)} \mathrm{E}[Y_{n+1}|x_{n+1},y,x,\rho_n,s_{n+1}]p^*(s_{n+1}|x,\rho_n,x_{n+1})\right) p^*(\rho_n|y,x,x_{n+1}). \tag{16}$$

The order restriction on the partitions now leads a simple covariate-dependent allocation scheme for the next subject

$$p^*(s_{n+1}|x,\rho_n,x_{n+1}) = \frac{p^*(\rho_{n+1}|x,x_{n+1})}{\sum_{s_{n+1}\in\mathcal{C}(\rho_n)}p^*(\rho_n,s_{n+1}|x,x_{n+1})} = \frac{p^*(\rho_{n+1}|x,x_{n+1})}{p^*(\rho_n|x,x_{n+1})}, \tag{17}$$

which can be computed from (14). In particular, we obtain that, conditionally on x,ρ_n,x_{n+1},

- If x_{n+1} is an end point (i.e. $x_{n+1}<x_{(1)}$ or $x_{n+1}>x_{(n)}$), the ordering constraint implies that there are two possible partitions of the $n+1$ data points. Suppose $x_{n+1}<x_{(1)}$, then either (i) the new data point is in the first cluster with probability proportional to $\frac{n_1}{n_1+1}$, or (ii) the new data point is in a new cluster with probability proportional to $\frac{\alpha}{k+1}$.

- If $x_{\pi_x(i)} < x_{n+1} < x_{\pi_x(i+1)}$ and $s_{\pi_x(i)} = s_{\pi_x(i+1)} = j$, the ordering constraint implies that there is one possible partition of the $n+1$ data points and new data point is in cluster j.

- If $x_{\pi_x(i)} < x_{n+1} < x_{\pi_x(i+1)}$ and $s_{\pi_x(i)} \neq s_{\pi_x(i+1)}$, the ordering constraint implies that there are three possible partitions of the $n+1$ data points. Either (i) the new data point is in the cluster j with probability proportional to $\frac{n_j}{n_j+1}$, (ii) the new data point is in the cluster $j+1$ with probability proportional to $\frac{n_{j+1}}{n_{j+1}+1}$, or (iii) the new data point is in a new cluster with probability proportional to $\frac{\alpha}{k+1}$.

As for the joint DPM model, $p^*(\rho_n|x, y, x_{n+1}) \neq p^*(\rho_n|x, y)$; yet notice that computations here are different because we do not require that x is random; thus we do not have a probabilistic model for x to use in computations. The resulting expression of the curve estimate is given in the following

Proposition 2. *If the random partition model is defined by (14), then the prediction of y_{n+1} given x_{n+1} and the data is*

$$\widehat{m}(x_{n+1}) = \sum_{\rho_n \in \mathcal{C}_n} \frac{1}{c} \tilde{m}(x_{n+1}; \rho_n)\; p^*(\rho_n|y, x),$$

where $c = ((\alpha + n)p(y|x, x_{n+1}))/((n+1)p(y|x))$ and

$$\tilde{m}(x_{n+1}; \rho_n) = \begin{cases} \frac{\alpha}{k+1}\beta_0'\underline{x}_{n+1} + \frac{n_1}{n_1+1}\widehat{\beta}_1'\underline{x}_{n+1} & \text{if } x_{n+1} < x_{\pi_x(1)}, \\ \frac{\alpha}{k+1}\beta_0'\underline{x}_{n+1} + \frac{n_k}{n_k+1}\widehat{\beta}_k'\underline{x}_{n+1} & \text{if } x_{n+1} > x_{\pi_x(n)}, \\ \frac{\alpha}{k+1}\beta_0'\underline{x}_{n+1} + \frac{n_j}{n_j+1}\widehat{\beta}_j'\underline{x}_{n+1} + \frac{n_{j+1}}{n_{j+1}+1}\widehat{\beta}_{j+1}'\underline{x}_{n+1} & \text{if } x_{\pi_x(i)} < x_{n+1} < x_{\pi_x(i+1)} \text{ and } s_{\pi_x(i)} = j, s_{\pi_x(i+1)} = j+1, \\ \frac{n_j}{n_j+1}\widehat{\beta}_j'\underline{x}_{n+1} & \text{if } x_{\pi_x(i)} < x_{n+1} < x_{\pi_x(i+1)} \text{ and } s_{\pi_x(i)} = j, s_{\pi_x(i+1)} = j. \end{cases}$$

Proof. First notice that the posterior of $[\rho_n|y, x, x_{n+1}]$ in (16) can be written in terms of the posterior of $[\rho_n|y, x]$, since

$$\begin{aligned} p^*(\rho_n|y, x, x_{n+1}) &= \frac{p^*(\rho_n|x, x_{n+1})}{p^*(\rho_n|x)} \frac{p^*(\rho_n|x)}{p^*(y|x, x_{n+1})} p(y|\rho_n, x) \\ &= \frac{p^*(\rho_n|x, x_{n+1})}{p^*(\rho_n|x)} \frac{p^*(y|x)}{p^*(y|x, x_{n+1})} p^*(\rho_n|y, x). \end{aligned}$$

Thus,

$$\widehat{m}(x_{n+1}) = \sum_{\rho_n \in \mathcal{C}_n} \left(\sum_{s_{n+1} \in \mathcal{C}(\rho_n)} \mathrm{E}[Y_{n+1}|x_{n+1}, y, x, \rho_n, s_{n+1}] \frac{p^*(\rho_{n+1}|x, x_{n+1})\, p^*(y|x)}{p^*(\rho_n|x)\, p^*(y|x, x_{n+1})} \right) p^*(\rho_n|y, x),$$

and using expression (14) to compute $p^*(\rho_{n+1}|x, x_{n+1})/p^*(\rho_n|x)$ or combining the allocation scheme (17) with expression (14) to compute $p^*(\rho_n|x, x_{n+1})/p^*(\rho_n|x)$, we obtain the result.

◇

Proposition 2 shows that, given the partition, the point prediction is an average of predictions based only on clusters close to x_{n+1} in the covariate space, where higher weight is given to neighbouring clusters with many individuals. Also, smaller α and larger k will give less weight to the prediction from a new cluster.

4 Computation

By enforcing an ordering constraint on the partition based on the covariate, we have reduced the number of possible partitions of n subjects into k groups from $S_{n,k}$, a Stirling number of the second kind, to $\left(\begin{smallmatrix} n-1 \\ k-1 \end{smallmatrix} \right)$; the first cluster must start with the first subject and there are $\left(\begin{smallmatrix} n-1 \\ k-1 \end{smallmatrix} \right)$ ways to choose where to start following $k-1$ clusters among $n-1$ remaining subjects. Thus, the constraint imposed reduces the total number of partitions from B_n to

$$\sum_{k=1}^{n} \left(\begin{smallmatrix} n-1 \\ k-1 \end{smallmatrix} \right) = 2^{n-1}.$$

However, for moderate to large n, this number is still large, and one needs to resort to MCMC methods to approximate $p^*(n_1, \ldots, n_k, k|y, x)$. To explore the space of partitions, we use the reversible jump MCMC algorithm as described in Fuentes-Garcia *et al.* (2010) and briefly described in the following paragraph.

At each iteration, one of two types of moves is proposed: a split, where a group of size bigger than one is divided into two, so that k is increased by 1, or a merge, where two neighbouring groups are combined, so that k is decreased by 1. Uniform distribution are used for both types of moves, so that

$$p^*(n_1, \ldots, n_{k+1}, k+1|n_1, \ldots, n_k, k) = \frac{1}{k_g(n_h - 1)},$$
$$p^*(n_1, \ldots, n_{k-1}, k-1|n_1, \ldots, n_k, k) = \frac{1}{k-1},$$

where for a split, h is the group selected to split and k_g is the number of groups of size larger than one. Letting $n^{(k)} = (n_1, \ldots, n_k)$, the acceptance probabilities for a split or merge, respectively, are

$$a(n^{(k+1)}, k+1|n^{(k)}, k) = \min\left\{1, \frac{p^*(n^{(k+1)}, k+1|x, y)}{p^*(n^{(k)}, k|x, y)} \frac{k_g(n_h - 1)}{k}\right\},$$
$$a(n^{(k-1)}, k-1|n^{(k)}, k) = \min\left\{1, \frac{p^*(n^{(k-1)}, k-1|x, y)}{p^*(n^{(k)}, k|x, y)} \frac{k-1}{(k-1)_g(n_{h_1} + n_{h_2} - 1)}\right\},$$

where for a merge, (h_1, h_2) are the two groups selected to merge and $(k-1)_g$ is the number of groups of size larger than one under the proposed merged partition. The proposed move is then accepted with its corresponding acceptance probability. Next, a shuffle of the current partition is performed, where two adjacent groups of size (n_{h_1}, n_{h_2}) are merged and then split into two groups of size $(n_{h_1^*}, n_{h_2^*})$. The shuffle is accepted with probability

$$a(n^{(k)*}, k|n^{(k)}, k) = \min\left\{1, \frac{p^*(n^{(k)*}, k|x, y)}{p^*(n^{(k)}, k|x, y)}\right\}.$$

For prediction, we use the estimate of $p(\rho_n|y, x)$ from the MCMC algorithm. We consider all (ρ_{n+1}) whose restriction to the observed n subjects is in the set of (ρ_n) with positive estimated posterior probability. For each $\rho_n^{(l)}$ visited in the chain, the local prediction, $\widehat{\beta}_j^{(l)'}\underline{x}_{n+1}$, and the non-normalized weight given in Proposition 2, denoted $w(x_{n+1}; \rho_n^{(l)})_j$, are computed for $j \in \mathcal{C}(\rho_n^{(l)})$. The prediction of y_{n+1} given x_{n+1} and the data can be estimated by

$$\widehat{m}(x_{n+1}) \approx \sum_{l=1}^{L} \sum_{j \in \mathcal{C}(\rho_n^{(l)})} \frac{1}{\widehat{c}}\, w(x_{n+1}; \rho_n^{(l)})_j\ \widehat{\beta}_j^{(l)'}\underline{x}_{n+1},$$

where

$$\widehat{c} = \sum_{l=1}^{L} \sum_{j \in \mathcal{C}(\rho_n^{(l)})} w(x_{n+1}; \rho_n^{(l)})_j.$$

Note that because we have greatly reduced the parameter space, we are able to sample the partition jointly as opposed to the DPM and joint DPM models which require sampling from the full conditional of cluster label for each subject. This results in much faster MCMC computations and better mixing.

5 Simulated data examples

To illustrate the issues related to the large number of partitions and the implications on predictive performance, we consider three simulated data examples. The results with the DPM model and joint DPM model which assign a prior on the full partition space are compared the proposed model whose support is restricted to a small, reasonable subset of the full partition space.

First, we study a simple example with a piecewise linear regression function and no error, so that the two clusters are clear. A set of $n = 37$ data points were generated according to the following formulae;

$$y_i|x_i = \begin{cases} -x_i/8 + 5 & \text{if } x_i \leq 6 \\ 2x_i - 12 & \text{if } x_i > 6 \end{cases}; \qquad x_i = 0, 0.25, 0.5, \ldots, 8.75, 9.$$

The hyper–parameters are specified as follows: $\alpha = 1$, $a = 2$, $b = 1/4$,

$$\beta_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \text{ and } C = \begin{bmatrix} 1/144 & 0 \\ 0 & 1/4 \end{bmatrix}.$$

For the joint DPM model, in all examples the local model for X is $F_X = \mathrm{N}(\mu, \sigma_x^2)$ and the base measure for X is the conjugate normal inverse-gamma, i.e. $\mu|\sigma_x^2 \sim \mathrm{N}(\mu_0, \sigma_x^2 c^{-1})$ and $\sigma_x^2 \sim \mathrm{IG}(a_x, b_x)$. The additional hyperparameters for the joint DPM model of example 1 are $a_x = 1$, $b_x = 1$, $\mu_0 = 4.5$, $c = 1/4$.

To illustrate the difficulties with nonlinear regression, a simple example with a quadratic regression function is considered. For $i = 1, \ldots, 50$,

$$Y_i|x_i \overset{\text{ind}}{\sim} \mathrm{N}(x_i^2, 1)\,; \qquad X_i \overset{\text{iid}}{\sim} \mathrm{U}(-5, 5).$$

The hyper–parameters are specified as follows: $\alpha = 1$, $a = 2$, $b = 1$,

$$\beta_0 = \begin{bmatrix} -12 \\ 0 \end{bmatrix} \text{ and } C = \begin{bmatrix} 1/50 & 0 \\ 0 & 1/25 \end{bmatrix}.$$

The additional hyperparameters for the joint DPM model are $a_x = 1$, $b_x = 1$, $\mu_0 = 0$, $c = 1/4$.

Finally, a more complicated example with $n = 100$ is generated according to

$$Y_i|x_i \overset{\text{ind}}{\sim} \mathrm{N}(x_i \sin x_i, 16^{-1})\,; \qquad X_i \overset{\text{iid}}{\sim} \mathrm{U}(-2\pi, 2\pi).$$

The hyper–parameters are specified as follows: $\alpha = 1$, $a = 2$, $b = 1/16$,

$$\beta_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \text{ and } C = \begin{bmatrix} 1/(72^2) & 0 \\ 0 & 1/144 \end{bmatrix}.$$

The additional hyperparameters for the joint DPM model are $a_x = 1$, $b_x = 1$, $\mu_0 = 0$, $c = 1/9$.

The MCMC scheme for the DPM model and joint DPM model (jDPM) is the Gibbs sampling method described in Neal (2000) (Algorithm 2). For the restricted DPM (rDPM) model, the algorithm described in Section 4 is used. All MCMC algorithms used 10,000 iterations with 1,000 burn in.

Example 1. We begin by analysing the posterior probability of the partition for the n observed subjects, since the prediction is computed based on those partitions with positive estimated probabilities. This first example demonstrates how inference for the random partition of the DPM and jDPM models can be (extremely) poor. Figure 1 summarizes the posterior of the partition by displaying the three partitions with the highest estimated probabilities for each of the models along with their corresponding probabilities.

The DPM model does not recognize the true partition. It gives the most weight, 0.3973, to the partition where the subject with a covariate of 8 is wrongly placed in the first cluster ($x_i \leq 6$). This occurs because more subjects are in the first cluster. Even though the correct partition has the second highest estimated probability, this value is only 0.0695.

The jDPM model is an improvement; with an estimated posterior probability of 0.5317 for the true partition, it does better at recognizing the clusters. However, the undesirable partition where the subject with a covariate of 8 is allocated to the first cluster is still present with the second highest estimated posterior probability of 0.0493.

With an estimated posterior probability of 0.9031 for the true partition, the rDPM model is by far the best at distinguishing the clusters.

The curve estimates at $x = 0.2$, 3.3, 5.9, 6.2, 6.3, 7.9, 8.1, 8.7 for the three models are shown in Figure 4. Apart from the subject with a covariate of 6.2, the cluster allocation of the new subjects is clear; those with covariates of (0.2, 3.3, 5.9) should be placed in the first cluster and those with covariates of (6.3, 7.9, 8.1, 8.7) should be placed in the second cluster. However, even conditionally on the true partition of the observed data, the DPM and jDPM models give positive weight to the allocation of these new subjects to the opposite cluster. This causes an unnecessary averaging of cluster-specific predictions across clusters that is evident in Figures 4a and 4b. For partitions other than the true one, the conditional prediction is necessarily worse.

By placing zero prior mass on undesirable partitions, we ensure that conditional prediction is just based on neighbouring clusters and the conditional predictions based on undesirable partitions have no impact. The prediction is greatly improved (Figure 4c).

As suggested by the referees, we also explored a similar example where the second cluster consists of subjects with covariates $x < 3$ or $x > 6$. In this case, by construction the rDPM model is not able to recover the true partition as values of the parameters cannot be shared across clusters. However, prediction is still improved as it is based only on neighboring clusters. We refer the interested reader to the supporting information, where this extension of example 1 is discussed.

Example 2. For the second example, the three partitions with the highest estimated probabilities for the three models are depicted in Figure 2.

In this example, the posterior mass for the DPM and jDPM models is spread out across many partitions. In particular, with 10,000 iterations, after discarding the first 1,000, a total of

9,946 partitions are visited by the chain for the DPM model and this number is 9,834 for the jDPM model. Moreover, the total mass of the top three partitions is only 0.0021 for the DPM model and is 0.0028 for the jDPM model. With a total of 1,044 partitions with positive estimated posterior probability and a total mass of 0.2345 for the top three partitions, the posterior mass for rDPM model is much less spread out.

The curve estimate for x from -4.5 to 4.5 by unit of 1 for the three models is displayed in Figure 5. The curve estimate for the DPM model does not even interpolate the data, and while poor curve fitting for this dataset was expected, the results in Figure 5a can appear very surprising. This is of course an extreme example, but it does demonstrate how dramatically poor the prediction can be for the DPM model when the true regression function is nonlinear, suggesting that the DPM model should be used with caution if there is any doubt in the linearity of regression function.

Prediction for the jDPM model (Figure 5b) is much better but is pulled down in some regions due to the influence of predictions based on clusters in other parts of the covariate space. The prediction of the rDPM model is close to the truth for all subjects except for the subject with a covariate of 0.5 due to lack of data in that area.

Example 3. For the last example, the unknown curve is rapidly changing and requires many clusters to capture it. The three partitions with the highest estimated probabilities for the three models are depicted in Figure 3.

This example demonstrates how dramatically spread out the posterior for the partition can be for the DPM and jDPM models. No partitions are visited more than once for both the DPM and jDPM models. Thus, all 10,000 partitions have the same estimated posterior probability, and Figures 3a and 3b display three of them. These partitions are composed of many clusters, with an average number of clusters of 15 for the DPM model and 13 for the jDPM model. Of the partitions displayed in Figures 3a and 3b, most contain undesirable features. Nevertheless, all these partitions are used for prediction.

For the rDPM model, on the other hand, the posterior mass is much less spread out. A total of 1,480 partitions have a positive estimated posterior probability. All partitions require at least six clusters, where the majority, 86%, of partitions have between 7 and 9 clusters.

Figure 6 displays the prediction for x from -2π to 2π by a unit of $\pi/8$. The DPM model again gives a linear prediction and thus, cannot capture the nonlinear regression function. For

the jDPM model, the prediction is not able to react to local changes in the derivative of the curve as well as the rDPM model because it is overly influenced by data in distant regions of the covariate space.

Insert Table 1 here.

We compared the empirical L_2 prediction error between the estimated prediction and the true prediction, defined by $(1/m \sum_{j=1}^{m} (\widehat{y}_{n+j,\text{est}} - \widehat{y}_{n+j,\text{true}})^2)^{1/2}$, in the three examples. The results are summarized in Table 1 . As expected from the above discussion, the rDPM model outperforms, and the jDPM gives better results than the DPM model.

6 Extension to binary response and real data application

In this section, we present an application to Alzheimer's disease, where the aim is estimation of the curve representing the probability of disease as a function of asymmetry in the hippocampus. As the response is binary, we also discuss a simple extension of the model developed in Section 3 to handle this scenario.

Alzheimer's disease (AD) is an irreversible, progressive brain disease that slowly destroys memory and thinking skills, and eventually even the ability to carry out the simplest tasks (ADEAR (2011)). Unfortunately, definite diagnosis is typically unavailable. Biomarkers based on neuroimages are becoming increasingly popular tools for diagnosis and monitoring disease progression of AD; and hippocampal volume is one of the most widely studied AD neuroimaging biomarkers, as the hippocampus is a relatively easy brain structure to identify and is known to be affected by the disease. As the disease progresses, brain tissue in the hippocampus deteriorates, and it is believed that this tissue loss occurs asymmetrically with some initial findings supporting this theory (Shi *et al.* (2009)). In this study, our aim is to further the understanding of the behavior tissue loss in the hippocampus for AD and provide support for the theoretical behavior of asymmetrical tissue loss.

The data used in this study was obtained from the Alzheimer's Disease Neuroimaging Initiative database (`adni.loni.ucla.edu`), which has collected around 5,000 images which are publicly accessible at UCLA's Laboratory of Neuroimaging. *The ADNI was launched in 2003 by the National Institute on Ageing (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies*

and non-profit organizations, as a $ 60 million, 5-year public- private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials. The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California-San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 adults, ages 55 to 90, to participate in the research, approximately 200 cognitively normal older individuals to be followed for 3 years, 400 people with MCI to be followed for 3 years and 200 people with early AD to be followed for 2 years. For up-to-date information, see `www.adni-info.org`.

Also available from the ADNI database are summaries of the neuroimages, including the volume of various brain structures, such as the hippocampus. To measure asymmetrical hippocampal tissue loss, we consider the ratio of the volume of the left to right hippocampus, which is computed from the structural magnetic resonance image performed at the first visit for 377 patients, of which 159 have been diagnosed with AD and 218 are cognitively normal (CN). We let $y = 1$ indicate a healthy subject, and x represent the ratio of the volume of the left to right hippocampus. Our aim is estimation of the curve

$$m(x_{n+1}) = \mathrm{E}[Y_{n+1}|x_{n+1}] = P(Y_{n+1} = 1|x_{n+1}).$$

We extend the model of Section 3 to handle a binary response by building on local probit models. First, suppose the observed response for subject i, y_i, is the indicator that the latent variable, y_i^*, is positive, i.e $y_i = I_{y_i^*>0}$. The model for the latent y_i^*'s is

$$Y_i^*|x_i, s_i = j, \beta^* \overset{\text{ind}}{\sim} \mathrm{N}(\beta_j^{*\prime}\underline{x}_i, 1),$$

where $\beta_j^* \overset{\text{iid}}{\sim} \mathrm{N}(\beta_0, C^{-1})$, for $j = 1, \ldots, k$, and the prior of the partition is given by the restricted random partition model in Section 3.

Simple calculations show that given the partition, the latent (y_i^*) are independent across

clusters and have multivariate normal distribution within cluster with parameters $\widehat{\underline{y}}_j^*$ and $\widehat{W}_j^{-1}$,

$$p(y^*|x,\rho_n) = \prod_{j=1}^{k}(2\pi)^{-n_j/2}\frac{|C|^{1/2}}{|C+X_j'X_j|^{1/2}}\exp\left(-\frac{1}{2}(\underline{y}_j^*-\widehat{\underline{y}}_j^*)'\widehat{W}_j(\underline{y}_j^*-\widehat{\underline{y}}_j^*)\right),$$

where $\widehat{\underline{y}}_j^*$ and $\widehat{W}_j$ are defined as in Section 2. Further conditioning on the response, we have that

$$p(y^*|x,y,\rho_n) \propto p(y^*|x,\rho_n) * \prod_{i=1}^{n}(I_{y_i^*>0})^{y_i}(I_{y_i^*\leq 0})^{1-y_i}.$$

Thus, given the partition and the data, the latent y_i^*'s are independent across clusters and have a truncated multivariate normal distribution within cluster with parameters $\widehat{\underline{y}}_j^*$ and $\widehat{W}_j^{-1}$ and regions defined by the observed responses.

The posterior of the partition given the data and the latent y_i^*'s is

$$p(\rho_n|x,y,y^*) \propto \frac{\alpha^k}{k!}\prod_{j=1}^{k}\frac{1}{n_j} * I_{s_{\pi_x(1)}\leq\ldots\leq s_{\pi_x(n)}} \quad \prod_{j=1}^{k}\frac{|C|^{1/2}}{|C+X_j'X_j|^{1/2}}\exp\left(-\frac{1}{2}(\underline{y}_j^*-\widehat{\underline{y}}_j^*)'\widehat{W}_j(\underline{y}_j^*-\widehat{\underline{y}}_j^*)\right).$$

Posterior samples of the partition can be obtained based on the MCMC algorithm discussed in Section 4 with an added step of sampling the latent y_i^*'s (see Damien & Walker (2001)).

Under the 0-1 loss function, the estimation of the regression curve amounts to determining

$$P(Y_{n+1}^*>0|x,y,x_{n+1}).$$

Given ρ_{n+1} and the latent y_i^*'s for the observed subjects, suppose the new subject is in cluster j, then Y_{n+1}^* is normally distributed with mean $\widehat{\beta}_j'\underline{x}_{n+1}$ and variance $\widehat{W}_{n+1,j}^{-1}$, as defined in Section 3.2. Thus,

$$P(Y_{n+1}^*>0|x,y,x_{n+1},y^*,\rho_{n+1}) = \Phi\left(\widehat{\beta}_j'\underline{x}_{n+1}\widehat{W}_{n+1,j}^{1/2}\right),$$

and the predictive probability of a success for the new subject is approximated by

$$P(Y_{n+1}=1|x_{n+1},y,x) \approx \sum_{l=1}^{L}\sum_{j\in\mathcal{C}(\rho_n^{(l)})}\frac{1}{\widehat{c}}\,w(x_{n+1};\rho_n^{(l)})_j\;\Phi\left(\widehat{\beta}_j^{(l)\prime}\underline{x}_{n+1}\widehat{W}_{n+1,j}^{1/2(l)}\right),$$

where

$$\widehat{c} = \sum_{l=1}^{L}\sum_{j\in\mathcal{C}(\rho_n^{(l)})} w(x_{n+1},\rho_n^{(l)})_j.$$

For the AD dataset, the hyperparameters are selected as

$$\beta_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad C^{-1} = \begin{bmatrix} 40 & 0 \\ 0 & 40 \end{bmatrix},$$

and $\alpha = 1$. The output of the MCMC algorithm with 20,000 iterations and 2,000 burn in was used to estimate the curve for new subjects with covariates of $x = 0.7$ to $x = 1.35$ by an interval of 0.01. Figure 7 displays the three partitions with the highest estimated posterior probability, and Figure 8 displays the estimated curve with 90% pointwise credible intervals computed from the output of the MCMC. The results show the presence of asymmetrical hippocampal volume in AD patients.

Under the 0-1 loss function, patients are classified as healthy if the estimated probability is greater than 0.5; new subjects whose left hippocampus is more than 11% smaller or more than 11% larger than than the right hippocampus are classified as sick. When the left hippocampus is more than 14% smaller than the right hippocampus the patient is classified as sick with at least 95% probability. This is comparable with the findings of Shi *et al.* (2009), who report a significant "left-less-than-right" hippocampal asymmetry pattern. However, our results also show that a "right-less-than left" hippocampal asymmetry pattern is present. In particular, the patient is classified as sick with at least 95% probability when the right hippocampus is more than 15% smaller than the left hippocampus.

7 Discussion

In this paper, we have provided a comparison of Bayesian nonparametric mixture models with constant versus covariate dependent weight functions for curve fitting, and identified a basic, but quite underestimated problem that is present in both models.

In terms of comparison, our results demonstrate an important drawback of the model with constant weight functions and linear mean functions; it is not robust to non-linearity in the regression function and can result in extremely poor prediction if non-linearity is present. This is due to the fact that inflexibility of the mean functions causes the clusters to be associated with regions of the covariate space. The local, cluster-specific predictions from different parts of the covariate space are averaged together, independently of x_{n+1}, resulting in poor prediction. To avoid this problem, single-p DDP models should use flexible mean functions that guarantee the curve described by the data can be captured by a single mean function. However, if the mean functions are too flexible, prediction will also suffer. On the other hand, we have shown that the model with covariate dependent weight functions results in improved prediction, due to the incorporation of prior knowledge of the partition structure based on the covariates.

However, for both models, problems arise due to the very large dimension of the partition space. In particular, the posterior puts too small a mass on desirable clusterings and too large a mass on undesirable partitions. Furthermore, an MCMC output may never even visit a partition with a desirable clustering. This occurs because it is not possible to manipulate the prior mass on partitions sufficiently, due to the extraordinarily large number of partitions and hence the microscopic probabilities involved. To address these issues, the prior knowledge on what are sensible configurations for the problem at hand needs to be introduced with extreme care. In fact, it is appropriate to rigidly restrict the support of the prior on the random partition to the set of sensible configurations, as this is the only sure way to guarantee prominence of desirable partitions in the posterior.

To make our point, we have focused on the particular case of simple regression, i.e. curve fitting, with a one-dimensional covariate, when it is essential to assume that clusters are based on covariate proximity. We have shown the importance of highlighting these clusters in the model by putting zero weight on the alternatives. The problems of not doing this, especially poor predictive performance, have been made evident through computations and a number of examples in the paper. For other applications, the type of clustering appropriate for the data or aim must be established, and once this is understood, undesirable partitions according to the notion of clustering established should be removed. We acknowledge that extensions to the case of multivariate regression are problematic, as there is no unique notion of ordering in higher dimensions. A general construction for the multivariate setting is provided in the supporting information, but a detailed extension is beyond the scope of this work and requires future research.

Acknowledgements.

We thank the referees and the AE and Editor for the constructive and careful comments.

Data used for the application in Section 6 of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (`adni.loni.ucla.edu`). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: `http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf`. Data collection and sharing for this application was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of

Health Grant U01 AG024904). We acknowledge the funding contributions of ADNI supporters (`adni-info.org/Scientists/ADNISponsors.aspx`).

S. Petrone was partially supported by grant 2008MK3AFZ of the Italian Ministry of University and Research, and by Bocconi University research grants.

Supporting information.

Additional information for this article is available online, including: an extension of Example 1 in Section 5 which is summarized by Figures S1 and S2; a discussion of extensions of the model to accommodate non-continuous or multivariate data; a complete list of ADNI sponsors; and all necessary **R** code.

References

ADEAR (2011). Alzheimer's disease education & referral center: Alzheimer's disease fact sheet. *NIH Publication* **11-6423**.

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* **2**, 1152–1174.

Barrientos, A. F., Jara, A. & Qunitana, F. A. (2012). On the support of MacEachern's dependent Dirichlet processes and extensions. *Bayesian Anal.* **7**, 277–310.

Blackwell, D. & MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes. *Ann. Statist.* **1**, 353–355.

Damien, P. & Walker, S. G. (2001). Sampling from truncated normal, beta, and gamma densities. *J. Comput. Graph. Statist.* **10**, 296–215.

De Iorio, M., Johnson, W. O., Müller, P. & Rosner, G. L. (2009). Bayesian nonparametric non-proportional hazards survival modelling. *Biometrics* **65**, 762–771.

De Iorio, M., Müller, P., Rosner, G. L. & MacEachern, S. N. (2004). An ANOVA model for dependent random measures. *J. Amer. Statist. Assoc.* **99**, 2205–215.

Denison, D. G. T., Holmes, C. C., Mallick, B. K. & Smith, A. F. M. (2002). *Bayesian methods for nonlinear classification and regression.* Wiley, Hoboken, New Jersey.

DiMatteo, I., Genovese, D. R. & Kass, R. E. (2001). Bayesian curve fitting with free-knot splines. *Biometrika* **88**, 1055–1071.

Dunson, D. B. & Park, J. H. (2008). Kernel stick-breaking processes. *Biometrika* **95**, 307–323.

Ewens, W. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology* **3**, 87–112.

Fan, Y., Dortet-Bernadet, J. L. & Sisson, S. A. (2010). A note on Bayesian curve fitting via auxiliary variables. *J. Comput. Graph. Statist.* **19**, 626–644.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1**, 209–230.

Fuentes-Garcia, R., Mena, R. H. & Walker, S. G. (2010). A probability for classification based on the mixture of Dirichlet process model. *J. Classification* **27**, 389–403.

Gelfand, A. E., Kottas, A. & MacEachern, S. N. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *J. Amer. Statist. Assoc.* , 1021–1035.

Griffin, J. E. & Steel, M. (2006). Order-based dependent Dirichlet processes. *J. Amer. Statist. Assoc.* **10**, 179–194.

Hannah, L., Blei, D. & Powell, W. (2011). Dirichlet process mixtures of generalized linear models. *J. Mach. Learn. Res.* **12**, 1923–1953.

Jara, A. (2007). Applied Bayesian non- and semi-parametric inference using DPpackage. *Rnews* **7**, 17–26.

Jara, A., Lesaffre, E., De Iorio, M. & Quintana, F. A. (2010). Bayesian semiparametric inference for multivariate doubly-interval-censored data. *Ann. Appl. Stat.* **4**, 2126–2149.

Kang, C. & Ghosal, S. (2009). Clusterwise regression using Dirichlet process mixtures. In *Advances in multivariate statistical methods* (ed A. Sengupta), 305–325. World Scientific Publishing Company, Singapore.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *Ann. Statist.* **12**, 351–357.

MacEachern, S. N. (1999). Dependent nonparametric processes. In *ASA proceedings of the section on Bayesian statistical science*, 50–55. American Statistical Association, Alexandria, VA.

MacEachern, S. N. (2000). Dependent Dirichlet processes. Tech. rep., Department of Statistics, Ohio State University.

Müller, P., Erkanli, A. & West, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika* **88**, 67–79.

Müller, P. & Quintana, F. (2010). Random partition models with regression on covariates. *J. Statist. Plann. Inference* **140**, 2801–2808.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Statist.* **9**, 249–265.

Norets, A. & Pelenis, J. (2012). Posterior consistency in conditional density estimation by covariate dependent mixtures. *Accepted by Econom. Theory* .

Park, J. H. & Dunson, D. B. (2010). Bayesian generalized product partition model. *Statist. Sinica* **20**, 1203–1226.

Pati, D., Dunson, D. B. & Tokdar, S. (2013). Posterior consistency in conditional distribution estimation. *J. Multivariate Anal.* **116**, 456–472.

Rasmussen, C. E. & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. MIT Press, Cambridge, MA.

Ren, L., Du, L., Dunson, D. B. & Carin, L. (2011). The logistic stick-breaking process. *J. Mach. Learn. Res.* **12**, 203–239.

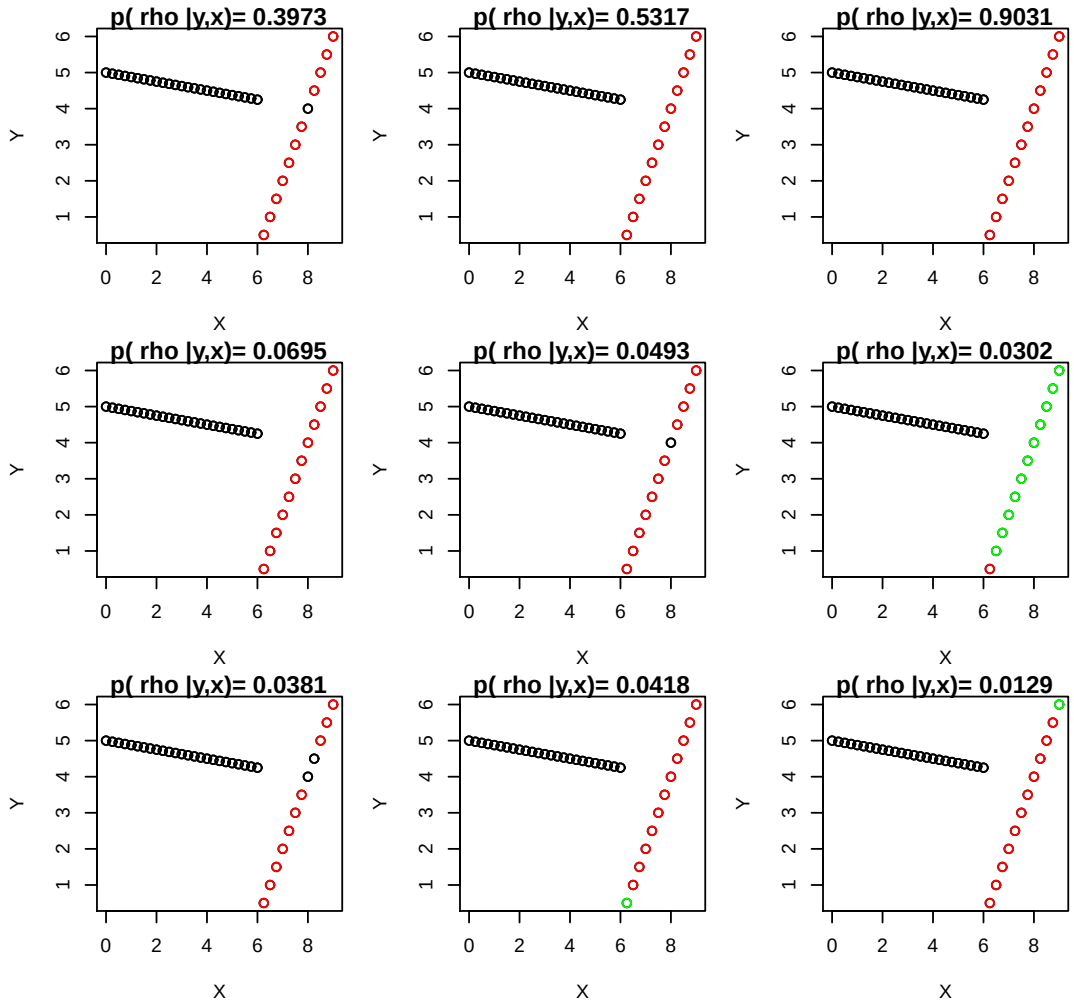
Rodriguez, A. & Dunson, D. B. (2011). Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Anal.* **6**, 145–178.

Shahbaba, B. & Neal, R. M. (2009). Nonlinear models using Dirichlet process mixtures. *J. Mach. Learn. Res.* **10**, 1829–1850.

Shi, F., Lui, B., Zhou, Y., Yu, C. & Jiang, T. (2009). Hippocampal volume and asymmetry in mild cognitive impairment and Alzheimer's disease: Meta-analyses of MRI studies. *Hippocampus* **19**, 1055–1064.

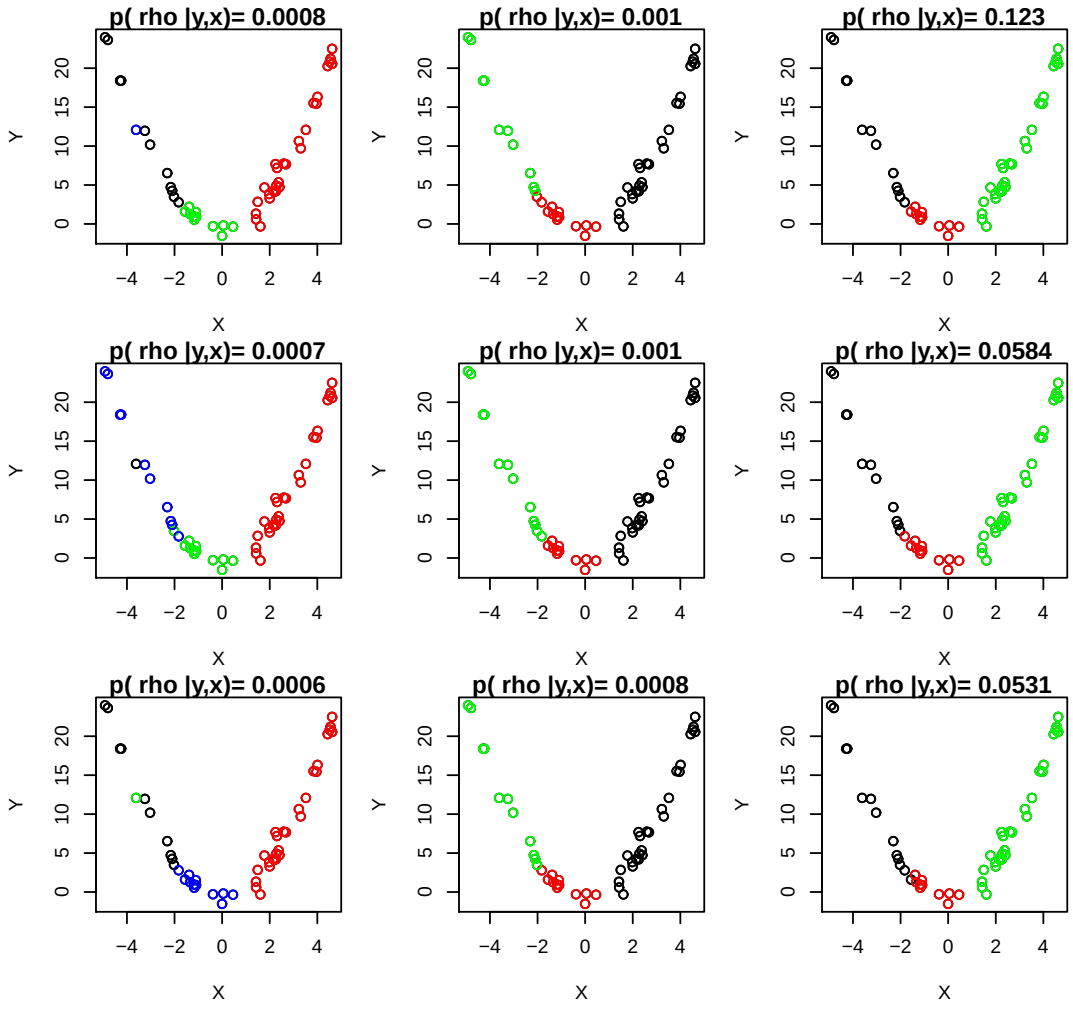
West, M., Müller, P. & Escobar, M. D. (1994). Hierarchical priors and mixture models, with applications in regression and density estimation. In *Aspects of uncertainty: A tribute to D.V. Lindley* (eds A. F. M. Smith & P. R. Freeman), 363–386. Wiley, Chichester.

Sara Wade, Computational and Biological Learning Laboratory, Department of Engineering, University of Cambridge, Trumpington Street, Cambridge CB2 1PZ, UK.
E-mail: sara.wade@eng.cam.ac.uk.



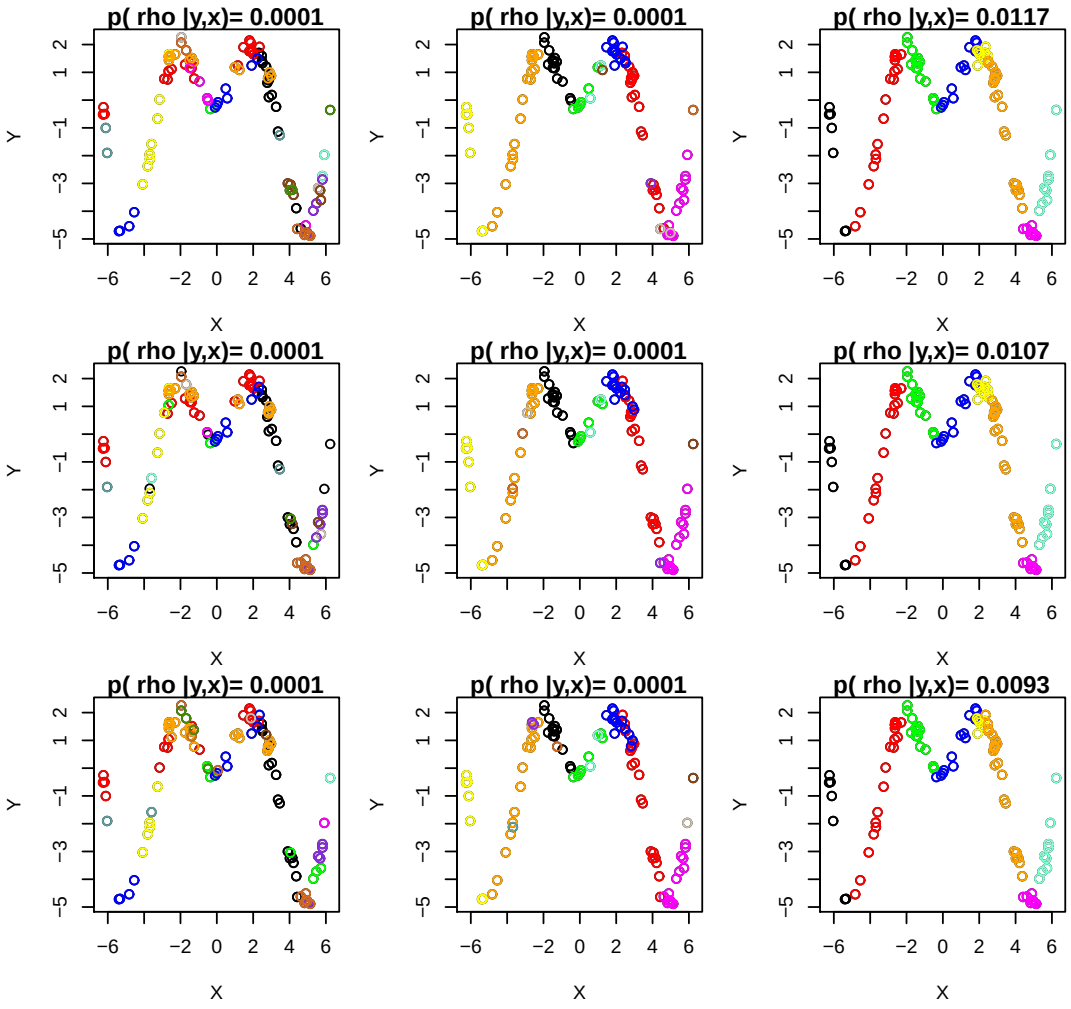
(a) DPM (b) joint DPM (c) restricted DPM

Figure 1: Simulated example 1. Data are generated with no error from two lines. The plots illustrate the posterior distribution on the unknown partition of the data obtained from a DPM, joint DPM and restricted DPM (by columns). The three partitions with the highest posterior probability (reported in the plot title) are shown by coloring the data according to cluster membership. The restricted DPM gives a much higher posterior probability (0.9031) to the correct partition.



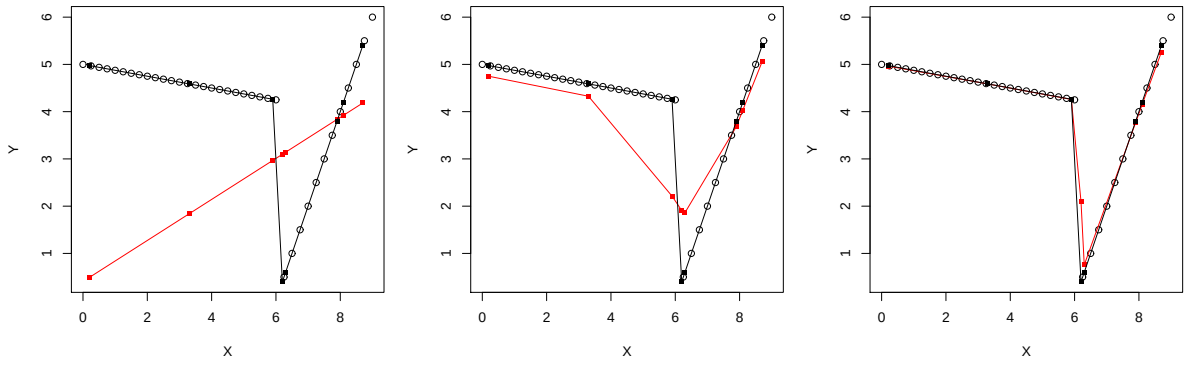
(a) DPM (b) joint DPM (c) restricted DPM

Figure 2: Simulated example 2. Data are generated as $Y_i|x_i \overset{\text{ind}}{\sim} N(x_i^2, 1)$. The DPM, jDPM and rDPM (by column) reconstruct the quadratic regression curve by locally selecting linear regressions corresponding to each cluster. The plots show the three partitions with the highest posterior probability, represented by coloring the data according to cluster membership. The very small values of the highest posterior probabilities (reported in the plot title) show that the posterior for the DPM and jDPM is very spread out.



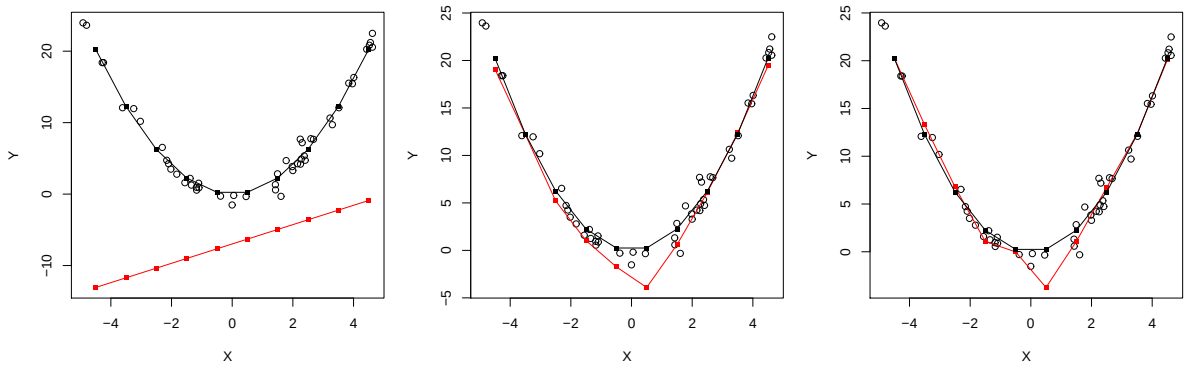
(a) DPM (b) joint DPM (c) restricted DPM

Figure 3: Simulated example 3. Data are generated as $Y_i|x_i \overset{\text{ind}}{\sim} \mathrm{N}(x_i \sin x_i, 1/16)$. The posterior distributions on the partition obtained from the DPM and jDPM are extremely spread out in this example; in fact, they are uniformly distributed over the $10,000$ partitions visited by the chain. The plots show three of these partitions for the DPM and jDPM (columns 1 and 2) and the three partitions with the highest posterior probability for the rDPM (column 3), by coloring the data according to cluster membership.



(a) DPM (b) joint DPM (c) restricted DPM

Figure 4: Simulated example 1. Plot of the curve estimate in red at $x = 3.3,5.9,6.2,6.3,7.9,8.1,10$, for the DPM, jDPM and rDPM, with the true curve in black and observed data in black circles.

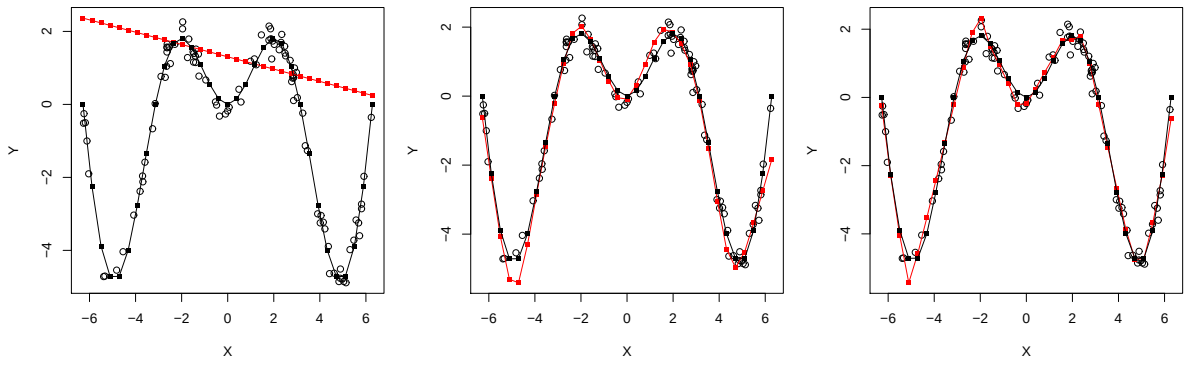


(a) DPM (b) joint DPM (c) restricted DPM

Figure 5: Simulated example 2. Plot of the curve estimate in red at a grid of new x values, together with the true curve in black and observed data in black circles. The poor result for the DPM is due to the fact that the curve estimate is an average of the linear regressions from all clusters (shown in Figure 3) independent of location of the new x value.

Table 1: Empirical L_2 prediction errors for the three simulated examples. The restricted DPM achieves the lowest error for all examples.

Example	DPM	joint DPM	restricted DPM
1	2.36	1.02	**0.60**
2	17.32	1.69	**1.42**
3	3.28	0.44	**0.26**



(a) DPM (b) joint DPM (c) restricted DPM

Figure 6: Simulated example 3. Plot of the curve estimate in red for a grid of new x values with the true curve in black and the observed data in black circles.

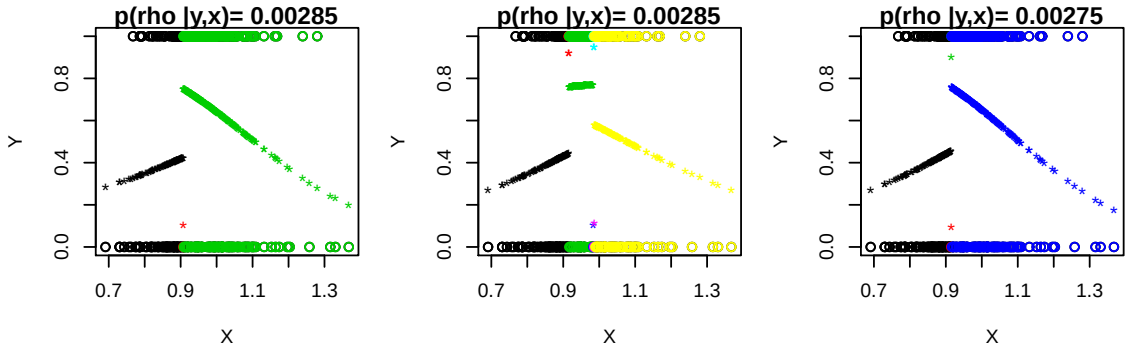


Figure 7: AD example. Plot of data colored by cluster membership for the three partitions with the highest estimated posterior probabilities (reported in the plot title). The plots include the within cluster estimated probit regression curve denoted with "*" and colored accordingly.

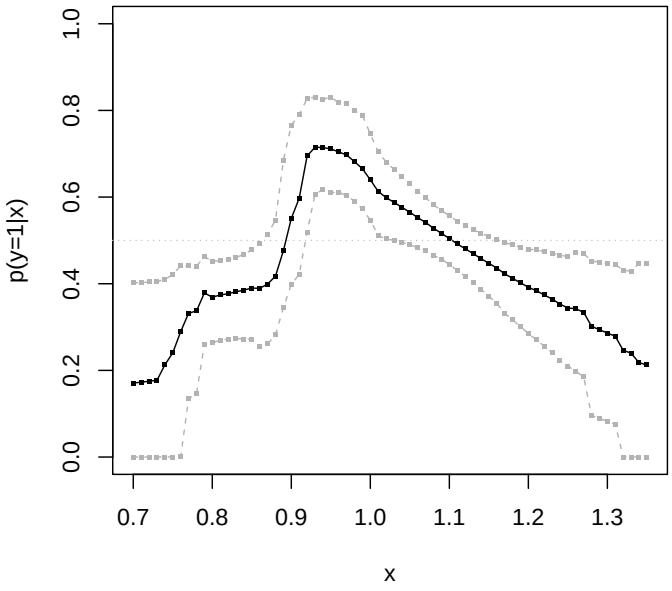


Figure 8: AD example. The estimated curve describing the probability of being healthy (in black) for left-to-right hippocampus ratios of 0.7 to 1.35 by 0.01 with 90% credible intervals (in gray).